Poster I-15

Modeling Microarray Data Using Bayesian Networks

Herskovits, Edward H.
University of Pennsylvania, Philadelphia, PA, USA

The small numbers of samples and large numbers of genes in microarray data sets preclude the application of conventional statistical methods; researchers have implemented analyses based on support-vector machines [1], cluster analysis [2-4], fuzzy logic [5], self-organizing maps [6]; perceptrons [7], and other statistical approaches [8-10]. Some of these approaches, particularly those based on clustering or on assumptions of multivariate Gaussian distributions for microarray data, are limited in the types of models they can generate, or equivalently, the types of gene-gene interactions they can capture from array data. In particular, clustering methods may not capture nonlinear multivariate interactions among genes, such as a model requiring that gene A be expressed only if genes B and C are expressed, and gene D is not expressed. To the extent that, for a given experimental condition and gene, gene expression follows a Gaussian distribution, we can model these data using a Gaussian mixture model (GMM) [11]. The utility of discretizing expression levels is indicated by researchers' tendencies to threshold expression levels (or their ratios) manually in an effort to determine which might be differentially expressed across sample classes (e.g., [12]). The principal advantage of representing expression levels using categorical variables is the existence of methods for capturing multivariate nonlinear relationships among these variables. The approach presented herein, called Bayesian Microarray Analysis (BMA), consists of converting expression levels into categorical variables [13]; representing these variables and (categorical) clinical variables as nodes in a Bayesian network [14]; and mining these categorical data for associations among the variables [15-17].

We tested these methods on the leukemia data from Golub et al. [18], and on the NCI data from Ross et al. [19], with the primary goal of histological tumor classification. For both data sets, BMA detected gene-histology associations that would be expected based on reports in the literature, as shown in Tables 1 and 2, respectively. For example, BMA found *Zyxin* to be strongly associated with the type of leukemia; in fact, this gene renders the Leukemia node conditionally independent of the remaining 7128 genes. Furthermore, as shown in Figure 1, even naïve Bayes classifiers with few genes demonstrate high classification accuracy.

Error! Not a valid link.

Table 1 Leukemia Genes (Partial List)

Zyxin
Phosphotyrosine independent ligand p62
Leptin receptor
C-myb
Cystatin A
Leukotriene C4 synthase (LTC4S)
CD33 antigen
Pentaxin-related gene
Adipsin
Azurocidin

Table 2 NCI Genes (Partial List)

Histology	Gene
Breast	P53
Breast	Efs1
Breast	Prolactin receptor
Breast	EDDR1
CNS	Sequence similar to pleiotrophin precursor
CNS	THY-1
Colon	ETS2
Colon	SLC9A1
Colon	Villin
Colon	GA733

References

1. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M, Jr., Haussler D. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 2000; 97(1):262–267.
2. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A 1998; 95(25):14863–14868.
3. Ben-Dor A, Shamir R, Yakhini Z. Clustering gene expression patterns. J Comput Biol 1999; 6(3–4):281–297.
4. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci USA 1999; 96(12):6745–6750.
5. Woolf PJ, Wang Y. A fuzzy logic approach to analyzing gene expression data. Physiol Genomics 2000; 3(1):9–15.
6. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 1999; 96(6):2907–2912.
7. Kim S, Dougherty ER, Chen Y, Sivakumar K, Meltzer P, Trent JM, Bittner M. Multivariate measurement of gene expression relationships. Genomics 2000; 67(2):201–209.
8. Raychaudhuri S, Stuart JM, Altman RB. Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 2000:455–466.
9. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 2000; 97(18):10101–10106.
10. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV. Fundamental patterns underlying gene expression profiles: simplicity from complexity. Proc Natl Acad Sci USA 2000; 97(15):8409–8414.
11. Titterington DM, Smith AF, Makov UE. Statistical Analysis of Finite Mixture Distributions. 1985, New York: John Wiley & Sons.
12. Amundson SA, Bittner M, Chen Y, Trent J, Meltzer P, Fornace AJ, Jr. Fluorescent cDNA microarray hybridization reveals complexity and heterogeneity of cellular genotoxic stress responses. 1999; 18(24):3666–3672.
13. Roberts SJ, Husmeier D, Rezek I, Penny W. Bayesian approaches to Gaussian mixture modeling. IEEE Trans Patt Anal Mach Intell 1998; 20(11):1133–1142.
14. Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. 1988, San Mateo, CA: Morgan Kaufmann Publishers.
15. Cooper GF, Herskovits EH. A Bayesian method for the induction of probabilistic networks from data. Machine Learning 1992; 9(4):309–347.
16. Herskovits EH. Computer-based probabilistic-network construction. In Medical Informatics. 1991, Stanford University.
17. Heckerman D. Bayesian networks for data mining. Data Min Knowl Disc 1997; 1(1):79–119.
18. Golub TR, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 1999; 286(5439):531–537.
19. Ross DT, et al. Systematic variation in gene expression patterns in human cancer cell lines. Nat Genet 2000; 24(3):227–235.